

Contingency Tables

42.2



Introduction

The practical application of statistics to engineering problems met in industry often concerns making decisions concerning probability distributions. For example you may be asked to decide whether a data set is approximately normal since much of the statistics you may apply makes this assumption. On occasions you may have to make such decisions given data concerning non-numeric variables in the form of a contingency table. Contingency tables are described in detail in this Workbook. This is one of the relatively rare occasions when hypothesis tests can be applied to non-numeric variables.



Prerequisites

Before starting this Section you should ...

- ① understand thoroughly what is meant by the term 'degrees of freedom'
- ② study the chi-squared distribution described in Workbook 40



Learning Outcomes

After completing this Section you should be able to ...

- ✓ understand what is meant by the term contingency table.
- ✓ perform hypothesis tests involving data given as a contingency table.

1. Contingency Tables

On occasions, it is possible that the members of a sample taken from a population can be classified by two different methods. Examples of this are:

- (a) articles produced by three machines running during two shifts on a production line;
- (b) the failure of electronic components and the position in which they are mounted in a machine;
- (c) the failure under compression testing of steel-alloy components and the rate of cooling applied during their production.

We can represent the information obtained by observation in such situations in a *contingency table*. By using the observed data to estimate expected data on the assumption that the classification methods are independent, we can use the chi-squared test to investigate the statistical independence (or otherwise) of the classification methods.

Consider the following contingency table with r rows and c columns. Such a table is referred to as an $r \times c$ contingency table.

	1	2	3	...	c	Row Totals
1	O_{11}	O_{12}	O_{13}	...	O_{1c}	R_1
2	O_{21}	O_{22}	O_{23}	...	O_{2c}	R_2
3	O_{31}	O_{32}	O_{33}	...	O_{3c}	R_3
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
r	O_{r1}	O_{r2}	O_{r3}	...	O_{rc}	R_r
Column Totals	C_1	C_2	C_3	...	C_c	N

Note that N is the total of the row totals and is the same as the total of the column totals, that is, N is the number of members of the sample taken from a population.

On the basis of the observed data we can estimate the expected frequency, say E_{ij} corresponding to the observed frequency O_{ij} . This is done as follows.

The probability that a randomly chosen element of the sample appears in row class i and column class j is given by p_{ij} where

$$p_{ij} = \frac{R_i}{N} \times \frac{C_j}{N}$$

Hence the required expected frequency is given by E_{ij} where

$$E_{ij} = N \times p_{ij} = N \times \frac{R_i}{N} \times \frac{C_j}{N} = \frac{R_i \times C_j}{N}$$

Using this formula repeatedly, we can calculate the expected frequencies corresponding to the observed frequencies and hence calculate a test statistic X^2 where

$$X^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

This formula tells you to calculate $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ for every cell in the contingency table and sum them.

It can be shown that, provided N is large, then the quantity

$$X^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

follows approximately a chi-squared distribution with $(r - 1) \times (c - 1)$ degrees of freedom. This number of degrees of freedom arises since each row has $r - 1$ independent entries and each column has $c - 1$ independent entries.

Note

The above statements are correct provided that we can calculate the expected frequencies without knowing the population parameters. If we have to estimate the population parameters, the number of degrees of freedom becomes $(r - 1) \times (c - 1) - m$ where m is the number of population parameters estimated. In the examples given here we shall not need to estimate the population parameters.

To complete the test procedure we note that the null hypothesis assumes class independence. For example, referring back to Example 2 given at the start of this Section, the null hypothesis would assume that the failure of electronic components and the position in which they are mounted in a machine are independent.

Should the test statistic exceed the critical value of χ^2 read off from table at (say) the 5% level of significance, we would reject the null hypothesis and conclude that a relationship of some kind exists between the classes.

It is worth noting that in some cases (such as the following worked example) one classification is chosen deliberately but the other is random while in other cases, both classifications are random. The same test applies in both cases.

Example In an experiment to determine the most advantageous position in a machine to mount an electronic component which may be prone to failure due to excessive heat build-up, 300 machines are tested with 100 randomly chosen examples of the component in each of 3 positions. The results obtained were as follows.

Position	1	2	3	Row Totals
Failure	40	30	50	120
Non-failure	60	70	50	180
Column Totals	100	100	100	300

Use a χ^2 -test at the 5% level of significance to determine whether component failure is related to mounting position.

Solution

The hypotheses are:

H_0 : component failure is independent of position,

H_1 : component failure is not independent of position

The expected frequencies are calculated as follows:

$$E_{11} = \frac{120 \times 100}{300} = 40, \quad E_{12} = \frac{120 \times 100}{300} = 40, \quad E_{13} = \frac{120 \times 100}{300} = 40$$

$$E_{21} = \frac{180 \times 100}{300} = 60, \quad E_{22} = \frac{180 \times 100}{300} = 60, \quad E_{23} = \frac{180 \times 100}{300} = 60$$

The test statistic is

$$\begin{aligned} X^2 &= \sum_{i=1}^3 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(40 - 40)^2}{40} + \frac{(30 - 40)^2}{40} + \frac{(50 - 40)^2}{40} + \frac{(60 - 60)^2}{60} + \frac{(70 - 60)^2}{60} + \frac{(50 - 60)^2}{60} \\ &= 0 + 2.5 + 2.5 + 0 + 1.67 + 1.67 = 8.34 \end{aligned}$$

and the number of degrees of freedom is $(r - 1) \times (c - 1) = (2 - 1) \times (3 - 1) = 2$ so that the critical value from tables is $\chi_{0.05,2}^2 = 5.99$.

Since $5.99 < 8.34$ we reject the null hypothesis and so we should conclude that there is a relationship between component failure and mounting position. Position 2 seems to be the most favourable and position 3 the least.



A factory runs three machines producing compression joints over a period of three daily shifts. In order to check whether there is any variability in the joints produced a random sample of joints produced is checked and the failures recorded. The results are given in the table below.

Shift	Machines			Row Totals
	1	2	3	
1	40	28	34	102
2	27	39	32	98
3	45	26	29	100
Column Totals	112	93	95	300

Use a χ^2 -test at the 5% level of significance to determine whether joint failure is related to the machine and the shift during which they were produced.

Your solution

The hypotheses are:

H_0 : joint failure is independent of shift,

H_1 : joint failure is not independent of shift

The expected frequencies are calculated as follows:

$$E_{11} = \frac{102 \times 112}{300} = 38.08, \quad E_{12} = \frac{102 \times 93}{300} = 31.62, \quad E_{13} = \frac{102 \times 95}{300} = 32.3$$

$$E_{21} = \frac{98 \times 112}{300} = 36.59, \quad E_{22} = \frac{98 \times 93}{300} = 30.38, \quad E_{23} = \frac{98 \times 95}{300} = 31.03$$

$$E_{31} = \frac{100 \times 112}{300} = 37.3, \quad E_{32} = \frac{100 \times 93}{300} = 31, \quad E_{33} = \frac{100 \times 95}{300} = 31.7$$

The test statistic is

$$X^2 = \sum_{i=1}^3 \sum_{j=1}^3 \frac{E_{ij}}{(O_{ij} - E_{ij})^2}$$

$$= \frac{38.08}{(40 - 38.08)^2} + \frac{31.62}{(28 - 31.62)^2} + \frac{32.3}{(34 - 32.3)^2} + \frac{36.59}{(27 - 36.59)^2} + \frac{30.38}{(39 - 30.38)^2}$$

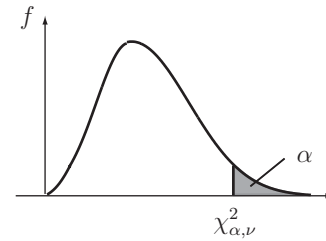
$$+ \frac{31.03}{(32 - 31.03)^2} + \frac{37.3}{(45 - 37.3)^2} + \frac{31}{(26 - 31)^2} + \frac{31.7}{(29 - 31.7)^2}$$

$$= 0.097 + 0.414 + 0.089 + 2.512 + 2.446 + 0.030 + 1.590 + 0.806 + 0.230 = 8.19$$

and the number of degrees of freedom is $(r - 1) \times (c - 1) = (3 - 1) \times (3 - 1) = 4$ so that the critical value from tables is $\chi_{0.05,4}^2 = 9.49$.

Since $8.19 < 9.49$ we do not have sufficient evidence to reject the null hypothesis and so we should conclude that there is no relationship between joint failure and the machine and shift on which the joints were produced.

Percentage Points $\chi^2_{\alpha, \nu}$ of the χ^2 distribution



α	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
ν											
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.01	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	31.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17